White Paper

The Long Tail of search - the value and the volume

November 2006

Key findings:

Using mathematics to determine the real life expectations of Long Tail and Deeper Buying

The effect Deeper Buying in searches can have on sales and search traffic volume

Generic searches compared to brand name searches and defining market share

The theory behind why brand name searches are more likely to result in a sale

www.searchlatitude.com

LATITUDE
LEADERS IN SEARCH

White Paper
Jon Myers
Search Director
Latitude

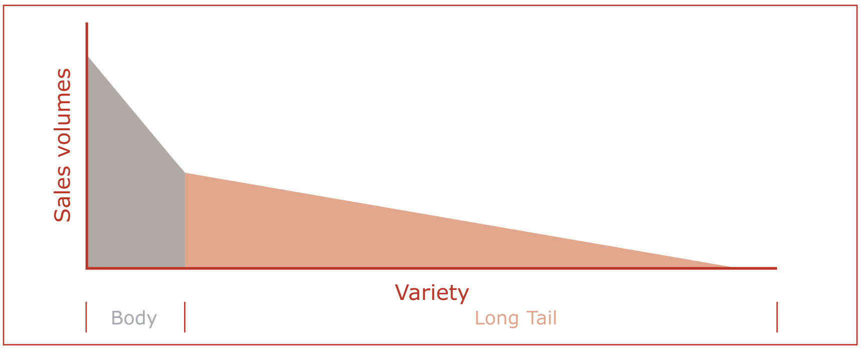
The Long Tail of search - the value and the volume

The expression The Long Tail was coined by Wired magazine editor-in-chief Chris Anderson in 2004, leading to a flurry of attention about the Long Tail phenomenon. But the term describes a feature of distribution graphs that has been recognised for decades.

In these graphs, a high frequency population — the head — is followed by a lower frequency population — the tail. Overall, the tail can outweigh the head. This sort of distribution is common not just in searches, but in overall daily occurrences. For instance, the words to and it are used frequently in the English language, but the less frequently used words represent the language as a whole.

Figure 1: Sales volumes v's variety

Figure 1



This same statistical distribution also occurs in studying search traffic patterns. Common keywords on which people search comprise the head, while obscure keywords form a Long Tail of 'one-hit wonders.' This phenomenon holds true for both paid and natural searches. However, for most retailers, their natural search traffic is made-up of almost entirely brand traffic. In reality, their Long Tail of unbranded traffic is quite short and raises many difficult questions, such as:

- How many search terms should drive traffic?
- How much traffic should any keyword drive?
- How does the volume of the Long Tail look?
- How does this compare to brand searches?
- What is the value of the Long Tail?

Rather than reviewing graphs about the Long Tail that can be found on thousands of Web sites, this paper will examine and analyse the actual mathematics of the volume and value of the Long Tail of Search. This will help to highlight how mathematics can be applied as a whole to the way individuals and search agencies use search engines such as Google, Yahoo and MSN on a daily basis.

What is the Long Tail?

The Long Tail hypothesis says that distribution is not a good reflection of buying habits. It holds that the unpopular items at the edges of the distribution when added up, contribute more than the popular items in the middle. That is, instead of being strictly bell-shaped, the distribution has long tails.

A good way of demonstrating this is using Zipf's law. For illustrative purposes, let's say an online dancing shoe retailer sells only three types of shoes: flamenco shoes, tap shoes and ballet shoes. And in total, they sell 183 pairs of shoes each day.

According to Zipf's law, the second most popular pair of shoes, tap shoes, are half as popular as the most popular pair of shoes, ballet shoes, while the third most popular pair of shoes, flamenco shoes, are one third as popular as the most popular pair (the ballet shoes).

Of the 183 pairs of shoes sold daily, ballet shoes account for 100 of these sales, tap shoes account for 50 (half as popular as ballet shoes) and flamenco shoes account for 33 (a third as popular as ballet shoes).

To calculate what fraction of total sales the most popular item (ballet shoes) are bringing in, it is 100 out of 183, expressed as a fraction of 100/183. Dividing 100 by 183 produces an answer of .546. Rounded out for expressive purposes, this tells us that ballet shoes account for approximately 55% of all items sold. So every item other than the most popular accounts for 45% of sales. Thus the tap shoes and the flamenco shoes represent 45% of all sales, while the ballet shoes represent 55% of all sales.

What did we really do to work this out? We simply took the ratio of the most popular item's sales to all items' sales in the following way:

Most popular item sales (1/1)*100
All item sales
(1/1)+(1/2)+(1/3)*100
Ratio =
[(1/1)*100]/{[(1/1)+(1/2)+(1/3)]*100} = (1/1)/[(1/1)+(1/2)+(1/3)]

The total number of sales is thus unimportant, as the 100 items cancel out top and bottom and we could have produced the same example using 1000 or 1,000,000. Regardless of the number used, we would still have found 55% of all sales were of the most popular item – the ballet shoes.

In this example, three items were being sold and we wanted to know how many sales the most popular item would account for.

In a more general example, n items might be sold and we want to know how much the most popular (k) items would account for. Then the above ratio would be:

Ratio =
[(1/1)+(1/2)+(1/3)+(1/4)+...+(1/k)]/[(1/1)+(1/2)+(1/3)+(1/4)+ ...+(1/n)]

As it gets larger, the ratio written above is close to log(k)/log(n). Just use the log key on any calculator. There are better approximations, but this is only a theory and it will make sense when applied to searches. For instance, here are some comparisons between summarising across the series and using log(k)/log(n)

k = 1000, n = 10000	Ratio from sums=0.765	ln(k)/ln(n)=0.75
k = 1000, n - 100000	Ratio from sums=0.619	ln(k)/ln(n)=0.60
k = 10000, n = 100000	Ratio from sums=0.810	ln(k)/ln(n)=0.80

Take this last example. What does the answer mean? It means that if I sell 100,000 different items, then the top 10,000 most popular will account for 80% of sales.

Now let's apply this to Search Terms and the deeper buying of keywords.

It is a close approximation, but let's assume there are 1,000,000 words in the English language. Depending on how many popular keywords we're interested in, then the fraction of searches they will account for are ln(k)/ln(1million)

Number of Search Terms by popularity = k

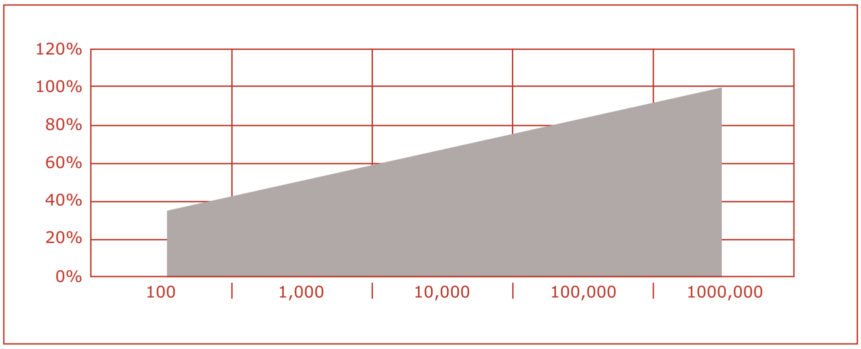
Number of words in the English language = n = 1,000,000

So, if we track the 100 most popular keywords, they will account for log(100)/log(1,000,000) =0.333 (33.3%) of all searches. In fact, assuming the total number of words in the search space (1,000,000 in the English language) never changes, we can approximately model how many of the most popular keywords in the language are necessary to cover a fraction of all searches:

Number of most popular keywords (k):	Fraction of all searches accounted for:
100	2/6=33.3%
1,000	3/6=50%
10,000	4/6=66.7%
100,000	5/6=83.3%
1million	6/6=100%

Figure 2: Volume of search against number of keywords

Figure 2



This, of course, is based on the Zipf's law model in which: “the second most popular keyword is half as popular as the first and the third most popular keyword is one third as popular as the first", and so on. This may or may not be true in the real world, but this exercise provides a good indication.

If the question is asked: “Under this model, how many of the least popular keywords in the English language account for the last 34% of all traffic?”, then we need to solve the equation [log(k)/log(n)] = 1-0.34 = 0.66.

We can already see that the answer for n=1,000,000 is pretty close to k=10,000 (meaning that the least popular 1,000,000-10,000= 990,000 keywords account for the last 34% of all traffic). If we want a formal solution:

For n=1,000,000, we get the answer: k = 9,120 (the most popular 9,120 keywords account for 66% of traffic and the least most popular 1,000,000-9,120=990,880 accounts for the remaining 34% of traffic).

If you only consider, say, n=10,000 words to be important words in the language for search purposes, then the same analysis would tell you that the most popular 437 words account for 66% of traffic and the least popular 9,563 words account for the remaining 34% of traffic.

What is the volume of the Long Tail?

The previous example effectively covers volume in terms of maths. But let's now apply this to an everyday search scenario.

The Automotive industry is a good sector to examine, due to the sheer number of keywords involved. Let's use the Audi brand as a hypothetical example and rank in order all the possible searches customers might make relevant to Audis:

Cars | New Car | Best Car | Fastest Car | German Cars...New Audi | New A3 | New A3 Audi...

The most popular searches are generic, making no mention of Audi. But the more targeted popular searches are branded or specific to Audi.

As a model, we will stick with Zipf's law, because it's the best way to interpret the Long Tail. But, as always, it assumes that the second search item is half as popular as the first and the third search term is a third as popular as the first, and so on.

1/1 | 1/2 | 1/3 | 1/4 | 1/5...
1/72 | 1/73 | 1/74...

So, for the Audi search terms above, in terms of popularity ranking:

This involves completely ignoring search terms that are relevant but are branded for other companies, and search terms that are completely irrelevant. Obviously, these terms will have a level of value to add volume, but let's focus on the top end and second tie of the tail.

Please note that only two groups of searchers are generally important to sellers. Customers who are making generic searches for products they sell and customers who are specifically looking for their brand. For this example, it is assumed the market share from irrelevant terms and customers looking for a competitor's brand are negligible.

Because of a Taylor Expansion of the series (1/1)+(1/2)....+(1/n), Zipf's law says that the fraction of searches in the first "k" search terms are well-approximated by ln(k)/ln(n).

I suggest assuming a break between generic and branded search terms to construct a mathematical model. So I'm going to say that the most popular "k" terms are unbranded or generic, and the next most popular n-k terms (where n is "all terms") are branded to Audi (more likely to result in a sale or deeper buying).

Under this model, if there are 100 relevant car-related search-terms, and the first 70 most popular are generic, but the next 30 (least popular) are branded, then the fraction of searches in branded terms is:

[ln(100)/ln(70)]-1 = 0.0839

(about 8.4% of searches are in terms branded to your company)

It again should be stressed that this is a model to illustrate the Long Tail. In the real world, it would be advisable to go out and source data to find which fraction of searches are generic and which are branded.

What is the value of the Long Tail?

Using the same idea for volume, let's now look at value.

What is this telling us about value?

The data suggests that there are (1/0.0839) = 11.9 times as many generic searches as branded searches.

So where does a company concentrate its resources?

In this instance, for it to be worthwhile pursuing the searches in the Long Tail, a customer only has to be 11.9 times (or more) as likely to buy based on a branded search than a generic search. If, as a total market, your customers who search for Audi-related terms are 12 times more likely to buy an Audi than those who search for "New Car" or other generic terms, then your

Figure 3

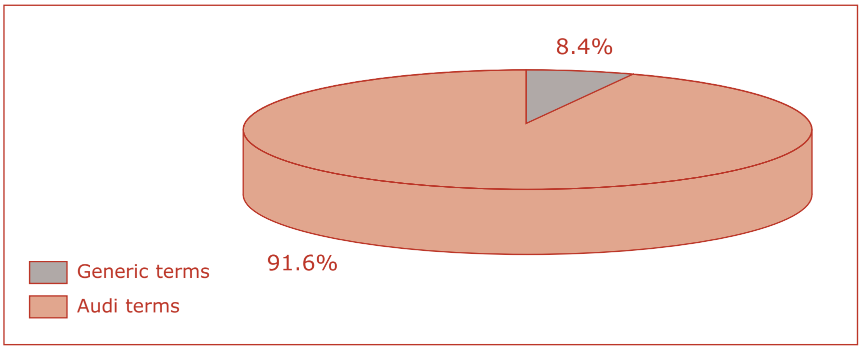


Figure 3: Purchase likelihood of generic v's brand terms

strategy is making good use of the Long Tail.

However, if, say there are only five major car companies and when generic searches like "New Car" are made, Audi ends up with 20% (equal market share) of those customers , Audi would simply not get 11.9 times 20% of the car market in deeper buying. So in this instance, their strategy should steer away from (or completely ignore) deeper buying. What would be happening here is that Audi would have more real-world impact, so it would not be worth their time to focus online advertising on deeper buying, when the generic customer is basically looking for an Audi anyway.

Audi would, of course, still pick up extra customers from deeper buying, but they should be asking themselves: is it worth focusing advertising revenue here, when I'll make most of my return from generic searches?

But for small companies, for example the hypothetical obscurecars.com (with Audi now becoming Big Cars Ltd.), deeper buying is a dream. Even though they may only have 0.1% in total market share, it would mean that 0.1% of users making generic searches buy their product. So if they can get 11.9*0.1%= 1.2% of users making branded searches to deeper buy their product, then they are massively increasing their selling potency over their typical market share.

Turning this on its head

You may ask: "given our market share and the typical number of search terms in our industry, how popular do our branded searches have to be to make deeper buying worth our while"?

This involves solving equations with logarithms. But we could extend this even further. What is probably happening in the real world is that there is a series in popularity that looks more like this:

GENERIC SEARCH TERMS...SEARCH TERMS SPECIFIC TO MY MORE POPULAR COMPETITORS' DEEPER BUYING...SEARCH TERMS SPECIFIC TO MY COMPANY'S DEEPER BUYING...SEARCH TERMS SPECIFIC TO MY LESS POPULAR COMPETITORS' DEEPER BUYING

So the fraction that you're getting in search terms branded to you is something like:

[ln(a)/ln(n)] + {[ln(b)/ln(n)] - [ln(a)/ln(n)]} + {1 -[ln(c)/ln(n)]}
Which simplifies to:
{[ln(c)/ln(n)] -[ln(b)/ln(n)]}

1+[ln(b)/ln(n)] -[ln(c)/ln(n)] OR, simpler again: {ln(n)/[ln(n)+ln(b)-ln(c)]}-1
and, using a property of log functions you may not know:
{ln(n)/[ln(n*b/c)]}-1
Where "a" was the rank in popularity of the last most popular generic search term
and "b" is the rank of the last of the more-popular-than-you search terms branded for your competitors

and "c" is the rank of the last search term branded for your company

This is the same equation as the more simplified example first provided. But the term "k", which was the edge between branded and generic ranking popularity, has now become the term "n*b/c", which is the edge between all terms that are generic to you, including more popular competitors, re-weighted over all relevant search terms, including less popular competitors.

So, if we assume that there are 100 relevant search terms (n=100), the first 50 are generic popular searches and the next 20 cover competitors with more popular branding than yours (b=70), then the next 20 are search terms relevant to your particular brand (c=90) and the last 10 of the 100 search terms are branded to companies less popular than yours.

The popularity rank of the effective edge between generic/more popular company terms and search terms branded for your company is (n*b/c=k=) 77.77. Then the fraction of searches covered by terms branded to you is (ln(n)/ln(77.77))-1=0.0577=5.8% of search terms. And there are (1-0.0577)/0.0577 =

16.3 times as many relevant searches that aren't related to your brand as are related to your brand.

Please note, if you ignore the terms for companies less popular than yours, then c=n=100 in the above example and you revert to the simple model with b=k (i.e. for the purposes of volume, you don't care at all whether search terms are generic or branded to a more popular company).

This is not true of value, because generic search terms should produce purchases at about your typical market share. However, search terms branded to other companies should produce zero purchases for you.

Taking value further

For extra value, you can then repeat the simpler example, and again talk about how much market share you need for deeper buying to have sufficient value.

This gets more complicated the more we consider value and your competitors' search terms.

As in the last formula:

"a" and "b" are the same

Also, in the first simple example

"k" was the rank in popularity of the last most popular generic search term.

Now, as we argued it, the relative value returned for you by generic search terms is just your typical real world market share. Let's call this "m". In the first example, your average market share in generic searches is just:

$$k*m/k = m$$

In the second example, because we assume that people buying into your competitors' branded terms never results in a sale for you (for the purpose of this, it may well result in sales), your average market share goes down to:

$$(a*m + [b-a]*0)/(b) = m*a/b$$

So, essentially, your market share drops by the fraction a/b, which in the example given is the fraction of all terms more popular than your terms that are in generic, rather than competitor-branded terms. In the example, this would be a/b = 50/70=0.7143

Now, let's run through the original numbers, where we constructed an example where a company had 0.1% market share and 11.9 times as many searches were made in generic terms than their branded terms. If they can get 1.2% of customers to deeper buy when they click through on their site, this is lucrative for them. Now, however, their market share is 0.7143*0.1% = 0.071%, which is even lower. It appears that the company only needs around 0.85% sales from deeper buying to make it worth their while.

How can it be that the better your competitors are doing, the less you need to work? It's because as your total market share shrinks, deeper buying looks more lucrative.

What is really being said here? The first thing to stress is that if we switch from the generic only to the "generic + more popular competitors" model, there is no real world effect on deeper buying. This is because normally we think of things relative to our market share in the real world, not pure market share in generic search + more popular competitor terms. The real world market share hasn't changed.

Conclusion

What can we conclude from this? What is happening in the example given is that your share of return on all searches that are more popular than your branded searches has dropped in value by 71.4%. This has two effects. First, you have to note that your return is in two components: generic or "fluke" buys and deeper buying:

Total = Generic + Deeper_Buying

(where Generic can equal Generic terms +more popular competitors branded terms)

So, if we include competitors who are beating you in popularity (if such competitors exist), the GENERIC return drops (say by my 71.4% example).

Total = (0.7142* Generic) + Deeper_Buying

However, deeper buying looks much more lucrative, because the ratio of Deeper Buying/Generic has grown.

Jon Myers
Jon is Search Director for Latitude and is based in the Cheshire office, UK.

If you have any comments or questions on this paper:

Jon.Myers@searchlatitude.com

So if things are compared to your real world market share, nothing has changed at all by including more popular competitors. However, if things are compared to a market share based on all search terms, including more popular competitors' branded terms, deeper buying looks more lucrative, because you're getting less return in the generic terms.

Many companies might not be too happy to hear that deeper buying looks like the correct strategy for them because it would mean all the "normal" buying was going to their competitors.

However, the important factor to consider is that the baseline market share should always be measured in terms of how many purchases your company gets in generic search terms – not in all search terms, including more popular companies' branded terms. Or it should be measured by some other real world brick-and-mortar store metric.

Of course, a genuine effect of including these more popular competitors' terms is that your potential for winning more market share in purely generic terms is probably small (because you can't make "new car" synonymous with "new Audi"). But there is potential for you to win popular search terms from competitors, because you can make "new Audi" a more popular search than "new BMW".

Were you to rerun the examples, and switch the ranks of your more popular competitors and your own company's (so that your company is now the most popular branded search), you would find you needed an even lower conversion rate on your deeper buying relative to your market share in generic terms. It's certainly still more lucrative to have more popular branded terms. This is a real effect on your volume.`

Summary

Provided that market share is always measured in terms of generic search terms, and your competitors' branded terms are ignored completely to get the baseline market share, then the mathematics of deeper buying is not affected by including your company's branded search terms as generic terms when measuring popularity in volume. However, you should always try to have the most popular non-generic terms, as this will maximise your total return from deeper buying by increasing your volume.

About Latitude

Latitude is the UK's leading independent Search Engine Marketing specialist, according to Companies House figures. Providing local and global Search Engine Marketing services in more than forty countries, as the market leader in SEM, Latitude's pay-per-click and search engine optimisation expertise have generated more than £500 million in online sales transactions in Europe alone.

Latitude is also one of the first UK agencies to provide pay-per-call and other innovative search options to our clients, who include Tesco Personal Finance, Totaljobs Group, Mansion and Betfair.

This year Latitude became a Media Momentum Award winner, featured in the Top 10 Sunday Times Tech Track fastest growing companies and CEO, Dylan Thwaites won the Ernst and Young National Entrepreneur of the Year – Technology Sector.

To receive future White Papers, please contact Matt Brocklehurst at the London office

Future Latitude White Papers

As Latitude is at the forefront of Search Engine Marketing in the UK, we continually monitor developments of interest to the industry and publish our findings on a regular basis.

For media queries or details on other White Papers please contact Matt Brocklehurst, Head of Marketing for Latitude on:
Email: matt.brocklehurst@searchlatitude.com

Web: http://www.searchlatitude.com

Tel: +44 (0) 20 7952 8000 (London office)

+44 (0) 1925 413 513 (Cheshire office)

Latitude Group Limited

Cheshire Office:
700 Mandarin Court
Warrington
WA1 1GG
Tel: +44 (0) 1925 413 513

London Office:
55 New Oxford Street
London
WC1 1BS
Tel: +44 (0) 20 7952 8000

www.searchlatitude.com

LATITUDE
LEADERS IN SEARCH